23. MORE HYPOTHESIS TESTING

The Logic Behind Hypothesis Testing

For simplicity, consider testing $H_0: \mu = \mu_0$ against the two-sided alternative $H_A: \mu \neq \mu_0$.

Even if $H_0$ is true (so that the expectation of $\overline{X}$ is $\mu_0$), $\overline{x}$ will probably not equal $\mu_0$ exactly.

Instead, we need to decide if the observed difference between $\overline{x}$ and $\mu_0$ can plausibly be accounted for by chance (i.e., by the natural variability of $\overline{X}$) or should be attributed to a systematic difference between the true and hypothesized means, $\mu$ and $\mu_0$.

If $H_0$ is true, then $Z$ is approximately standard normal, and will very rarely lie outside the interval $(-z_{\alpha/2}, z_{\alpha/2})$. 
But if \( \mu \neq \mu_0 \) then the distribution of \( Z \) will have a nonzero mean, with the same sign as \( \mu - \mu_0 \), and it would not be so unusual to find \( z \) in the rejection region.

So if for our given data we find that \( z \) is in the rejection region, there are only two possibilities:

- **EITHER** \( H_0 \) is true, in which case the observed value of \( z \) must be just a “fluke”, or rare event, due simply to the natural variability of \( \bar{X} \); (This “false alarm” scenario is not impossible, although it is somewhat implausible, especially if \( \alpha \) is small),

- **OR ELSE** \( H_0 \) must be false.
Here, a reasonable person would conclude that there is sufficient evidence to reject $H_0$.

The situation is analogous to having an alarm which almost never goes off falsely, but which is now ringing.

It is more plausible that the largeness of $|z|$ is caused by some systematic effect (i.e., that $\mu \neq \mu_0$), rather than by the natural variability of a standard normal. Thus, we reject $H_0$. 
To improve your understanding of the discussion above, consider [R Demo: Power].

This gives graphs illustrating the power of the test. The power is the probability of rejecting the null hypothesis. It depends on the value of the population mean.

The R demo uses the situation in Example 1 of the previous handout (Quarter Pounders), which was a left-tailed test, with hypothesized mean $= 0.25$, $\sigma = 0.035$, $n=50$. The values of the population mean $\mu$ are 0.25, 0.24, 0.23, 0.22.
Statistical Significance And
The Meaning Of $\alpha$

• If $H_0$ is rejected, we say that the results are **statistically significant** at level $\alpha$.

In this case, we have proven that $H_A$ is true, beyond a reasonable doubt (but not beyond all doubt).

Note that $\alpha$ *is not* the probability that $H_0$ is true, since there is nothing random about $H_0$.

Instead, $\alpha$ represents the false alarm rate (Type I error rate) of the test, i.e., the proportion of the time that a test *of this kind* would reject $H_0$ if $H_0$ were in fact true.
A finding of statistical significance does not provide absolute proof that $H_0$ is false.

We may be committing a Type I error (i.e., we may have a false alarm).

To make matters worse, we may never find out whether we made a mistake by rejecting $H_0$.

We do know, however, that if $H_0$ were true, then false alarms would be unlikely to occur: they would have probability $\alpha$.

• If $H_0$ is not rejected, then we say that the results are not statistically significant at level $\alpha$. 
The terminology often used here is that $H_0$ is “accepted”, but this should be avoided, since our inability to find sufficient evidence to reject $H_0$ does not in any way demonstrate that $H_0$ is true. (By analogy, the acquittal of a defendant on murder charges obviously does not constitute proof of innocence.)
Tests For $\mu$ When $\sigma$ Is Unknown

When $\sigma$ is unknown, we estimate it by the sample standard deviation, $s$.

The test statistic to use in this case is $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$.

The $t$-statistic measures how far the sample mean is from the hypothesized population mean, in units of estimated standard errors.

If the population is normal and $H_0$ is true, then $t$ has a Student’s $t$ distribution with $n - 1$ degrees of freedom.
The criteria for a level $\alpha$ test are:

<table>
<thead>
<tr>
<th>$H_A$</th>
<th>Rejection Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu \neq \mu_0$</td>
<td>$</td>
</tr>
<tr>
<td>$\mu &lt; \mu_0$</td>
<td>$t &lt; -t_{\alpha}$</td>
</tr>
<tr>
<td>$\mu &gt; \mu_0$</td>
<td>$t &gt; t_{\alpha}$</td>
</tr>
</tbody>
</table>

This test is commonly referred to as the $t$-test.

Values of $t_{\alpha}$ can be found in Table 6, using df = $n - 1$.

As df gets larger, $t_{\alpha}$ becomes smaller.

For df $\geq 29$, $t_{\alpha}$ and $z_{\alpha}$ are reasonably close. (We use the same cutoff for “large sample sizes” as we did in constructing confidence intervals. See the discussion given there.)
Before applying the $t$-test, it is wise to check a histogram of the data for approximate normality. Although it is safe to apply the $t$-test even if the data contain outliers, the actual level (false alarm rate) of the test will typically be somewhat smaller than $\alpha$ in this case.

A more serious problem is that the probability of a Type II error will typically be larger, so the test has a harder time detecting that $H_A$ is true, than in the normal case.
Eg 1: A 2011 Volkswagen Jetta was tested for NO\textsubscript{x} (nitrogen oxide) emissions in two runs on an urban route in Los Angeles, CA. The Environmental Protection Agency limit for this pollutant is .04 g/km. The average emission was .989, with a sample standard deviation of .114. Is there evidence that this car violated the EPA limit, at the 1% level of significance? Assume that NO\textsubscript{x} emissions are normally distributed.

Sol: Let \( \mu \) represent the expected NO\textsubscript{x} emissions for this car and route. We want to test \( H_0: \mu = .04 \) versus \( H_A: \mu > .04 \). Thus, \( H_0 \) states that the VW Jetta did not violate the EPA limit, while \( H_A \) states that it did.
Eg 2: The manager of a credit card company claims that the mean time to settle disputed charges is 30 days. A regulator is worried that the manager’s claim is too optimistic. The regulator examines a random sample of 15 disputed charges, and finds a mean time to settlement of 35.9 days, with a sample standard deviation of 10.2 days. Is there evidence at the 5% level of significance to doubt the manager’s claim, assuming that the time to settle disputes is normally distributed?

Sol: Here, we test $H_0: \mu = 30$ versus $H_A: \mu > 30$, where $\mu$ is the population mean (that is, expected) time to settlement for all disputed charges.

Eg 3: The Lucky Coin Demo.